Multimedia Quality assessment

Fatih Porikli, Al Bovik, Chris Plack, Alregib Ghassan, Joyce Farrell, Patrick Le Callet, Quan Huynh-Thu, Sebastian Möller, Stefan Winkler

Fatih Porikli, Al Bovik, Chris Plack, Ghassan AlRegib, Joyce Farrell, Patrick Le Callet, Quan Huynh-Thu, Sebastian Möller, and Stefan Winkler

Multimedia Quality Assessment

This *IEEE Signal Processing Magazine* forum discusses the latest advances and challenges in multimedia quality assessment. The forum members bring their expert insights into issues such as perceptual models and quality measures for future applications such as three-dimensional (3-D) videos and interactivity media. The invited forum members are Al Bovik (University of Texas), Chris Plack (University of Manchester), Ghassan AlRegib (Georgia Institute of Technology), Joyce Farrell (Stanford University), Patrick Le Callet (University de Nantes), Quan Huynh-Thu (Technicolor), Sebastian Möller (Deutsche Telekom Labs, TU Berlin), and Stefan Winkler (Advanced Digital Sciences Center). The moderator of this forum is Dr. Fatih Porikli (MERL, Cambridge).

Our readers may agree or disagree with the ideas discussed next. In either case, we invite you to share your comments with us by e-mailing fatih@merl.com or spm.columns.forums@gmail.com.

Moderator: Let's start our discussion. Here is the first question on philosophical aspects: How would you define "objective quality" and "perceptual quality?" What are the attributes that characterize the picture (signal) quality?

Quan Huynh-Thu: I think that, generally speaking, "quality" is fundamentally the result of a human (and therefore subjective) judgment based on several criteria. Some of these criteria are measurable as they can be based on intrinsic information of the signal to judge, while other criteria are the result of cognitive processes integrating information beyond just the signal that is being judged (e.g., context, experience). So in a sense, quality is always perceptual. Now, the distinction I see concerns the algorithms (or metrics) that are designed to compute a measure of quality. I usually make the following distinction: "objective quality" is the value calculated by a computational model, whatever that model is. In other words, computational models produce an objective quality as a prediction of (subjective) quality. The simplest one, which is still widely used in the literature, is the peak signal-to-noise ratio (PSNR). However, such a model is known for not representing faithfully the human visual response, e.g., masking processes. I would then use "perceptual quality" to either represent subjective quality itself (i.e., as judged by human beings) or represent the value calculated by a computational model when its prediction is closely correlated with human judgment. Now, obviously, the difficulty and grey area is in how closely this should be. I guess I don't have a simple answer to this. Experts use different kinds of statistical analyses to quantify and characterize the performance of objective quality assessment models, but there is no defined or agreed simple threshold to define whether a model computes an objective quality that correlates enough with subjective quality so to be considered as perceptual quality. This is where it becomes a bit philosophical.

As expressed previously, I think that some of the attributes that characterize the picture quality can be objectively described from the extraction of features in the signal. The particular features that need to be measured depend highly on the type of signal to judge (e.g., picture, video, voice, and audio) but also on the context or types of processing/degradations that will impact the quality. Generally speaking, subjective experiments can be conducted to identify the particular attributes that are highly correlated with the human judgment of quality in the particular context of interest. Then, based on that knowledge, algorithms integrating the computation of these features can be developed.

Sebastian Möller: I think that there is no such thing like "objective quality." In my opinion, quality is the result of a perception and a judgment process, during which the perceiving human compares the features of the perceptual event, which is commonly provoked by the physical event, i.e., the signal reaching her perception organs, to the features of some internal reference. This reference reflects the expectations, experience, but also the transient internal states like motivation, emotions, and so on. Quality is thus inherently "subjective," determined by the reference of the perceiving human. It "happens" in a particular judgment situation. Still, scientifically speaking, we need to make the perception and judgment process as "objective" as possible, thus independent of the experimenter who performs the measurement.

If we want to find out about the attributes of quality, we need to identify the underlying features of the perceptual event. When judging transmitted speech under laboratory conditions, multidimensional scaling experiments as well as semantic differential techniques suggest that there are four to seven features, including sound quality, continuity, (absence of) noisiness, and loudness, to name just a few. However, additional features might arise if one digs further into one of these dimensions, or if experiments are carried out under more

Digital Object Identifier 10.1109/MSP.2011.942341
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE

realistic conditions. In this case, qualitative methods are a good starting point to identify relevant features.

Chris Plack: I disagree with this. Quality can be defined (by the Merriam Webster dictionary) as "an inherent feature" or "a property," or "a degree of excellence." Both of these can be objective. Clearly, when comparing systems, someone has to decide which measurable qualities are important to excellence, but that doesn't make the qualities themselves subjective.

If it is measurable using a physical instrument, then it is an objective quality. If measurement requires a response from a human observer, then it is a perceptual quality. For example, an image with a higher spatial resolution might be considered of higher quality, and hence system quality can be compared in terms of the objective quality of resolution. On the other hand, we might also ask participants to make a subjective rating of the "crispness" of an image, which is a perceptual quality.

Sebastian Möller: For me, the spatial resolution would be a metric for performance. The same holds for temporal resolution, bandwidth, etc. All these are entities that can easily be measured with a physical instrument. All may also have an influence on perceived quality, but they would not be qualities in my point of view—I would reserve this word for the perceptual entity. Having said this, it is also obvious that I disagree to quality being an "inherent feature" or a "property" (even if common usage of the term suggests so)—you could say that spatial resolution is an inherent feature, but it has to be judged by a human before being linked to quality.

Quan Huynh-Thu: I would tend to agree with Sebastian concerning the idea that quality is the result of a human judgment, as I wrote in my first postanswer. One of the reasons is that the concept of signal quality involves a reference point, which is mostly subjective.

Spatial resolution, temporal resolution, and bandwidth are all objectively measurable features but not qualities per se. In the example Chris gives, a higher spatial resolution image may be judged with higher quality but that is not necessarily so. It also depends on the other attributes/features. For example, at a given bit rate, encoding an image/video at resolution R1 versus resolution R2 with R1 > R2 will not necessarily produce a higher perceived quality; it all depends on the bit rate used and on the values of R1 and R2. At low bit rates, it is possible that an image/video at lower resolution will be judged as being of higher quality but, at very high bit rates, the reverse judgment of quality will likely happen.

Today, in Brittany, the weather temperature is 18 °C and rainy, although it is supposedly summer. Yesterday, it was 14 °C … so usual language could say that the weather is of better quality today but truly the only thing you one can say is that today is warmer. Yesterday was colder but sunnier, so but in my opinion the weather was of better quality yesterday because the sun has a bigger effect on me than temperature, but a Breton (who doesn't care if it rains 365 days a year and will complain that it is too hot when it's over 20°) may say or think otherwise as his/her reference is different.

Chris Plack: It depends on definition, of course. However if you define quality as a property or feature of something (which is one of the most common definitions), then clearly quality can be measured objectively.

OK, in the example I gave, if you fix the bit rate, then perceptual quality may not be predictable from resolution. However this is nit-picking in my view. In some cases, there may be an interaction between objective qualities to determine perceptual quality, but that doesn't deny the fact that there are objective qualities. Or that quality can be measured objectively. Would a better example be bit rate itself?

Patrick Le Callet: With increasing interests of the field witnessed by an explosion of literature, the definition of those concepts deserves clarification, as some meanings have been progressively lost. At this point, it is worth including also the concept of "subjective quality." Initially, "subjective" and "objective" have been introduced to differentiate two different ways to assess quality. Subjective quality assessment means measuring quality with humans, requiring experiments to get their opinions. Objective quality assessment refers to a metric that provides numbers in a repeatable manner. It covers a simple case of a signal attribute as a complex algorithm that could mimic human perception. Perceptual quality is related to the quality perceived by humans, involving sensation, perception, cognition of and context of use.

Coming back to quality assessment, by definition, subjective quality assessment should be able to capture perceptual quality. This is not necessarily the case for objective quality assessment. Getting some numbers computable in a repeatable manner is what is needed in many fields and is practically more useful. An objective quality assessment should provide this; nevertheless, the question remains open as an objective quality assessment tool is not necessarily able to reflect the perceptual quality in all contexts.

Attributes that characterize signal quality are numerous; their perceptual values are mainly context dependent. (I support Sebastian's point of view regarding this part.)

Ghassan AlRegib: As mentioned in the discussions so far, depending on the usage and the application (or context), one could define these two quality measure differently. In some cases, one might need an objective quality, which is obtained via computational models. In other cases, one needs the true perceived quality. In the former, I am thinking more in terms of applications where a "machine" needs to have some number to describe the quality of a signal. In the latter, the human wants to "judge" a signal. Of course, there has been a tremendous effort in coming up with models that generate an "objective quality" that simulates what a human brain does when judging such a signal. I tend to believe that in subjective quality, humans use their background experiences, perhaps some references, and many other factors while judging the quality of a signal. They do not tend to come up with a quality that fits a certain application. On the contrary, an "objective quality" measure or a metric has a target application (e.g., compression, streaming, etc.) and tries to come up with a number to be used in the target application. I think this viewpoint is in line with Sebastian's statements on "inherent property" and how they are linked to quality.

What I am saying is that subjective quality and objective quality are two different things; it is not the measure that is subjective or objective, it is the "quality" that is either subjective or objective. They of course differ in the mechanism but they also differ in the constraints, the factors, and the usage. There are some attributes in the signal and there are attributes in the application and the usage of the quality that affect the choice of objective versus subjective quality measures. This becomes clearer when we developed and tried to measure the quality in 3-D video and in human interaction applications. If you give a person a job to perform (e.g., tie a shoelace) in virtual reality using state-of-the-art interactive devices (e.g., data gloves) and assign an evaluator to watch the virtual action, how would the quality assigned by the evaluator differ from an objective quality measure? It is very complex signal and even to define a "good" shoe lace tying is a very tough task to start with. Similarly, a single monocular cue such as color or texture may have some impact on the quality of a two-dimensional (2-D) video, but they have catastrophic impact on the quality of a 3-D video. Nevertheless, humans may not notice such implications in a 3-D video, while they notice it clearly in a 2-D video. Back to the example of "high resolution means higher quality." In our lab, we ran many experiments with several subjects over a number of displays (a couple of years ago) to understand the impact of resolution, frame rate, texture content, etc. on the perceived quality of ultrahigh-definition (HD) videos as well as 3-D videos. High resolution gave "good experience" for a short period of time but ultimately the quality of the content (color, sharpness, texture etc.) decided on the perceived quality of videos. Limited resolution means, in my opinion, that we can hide several artifacts from human eyes. This takes me back to the discussions on how inherent features are linked to quality but quality is not an inherent feature or a property.

Al Bovik: Since the aim of an objective quality model is to predict, as accurately as possible, perceptual quality, then I will take the position that they are the same thing, in the sense that the goal is that they be the same thing. Naturally, we are not there yet; objective models deliver numbers, and objective experiments do the same, whereas visual quality is something ineffable, of infinite dimensions, and varying with situation, context, space, and time.

Yet we are on the right track toward solving these problems in psychology and engineering. As several others have expressed, perceptual modeling is a key ingredient to success. Low-level models of cortical and extra-cortical processing, as used in the past by Beau Watson, Stefan Winkler, and many others enable us to define perceptually meaningful attributes of visual signals. Similarly, statistical models of the naturalness of visual signals, provably regular over very large, generic, and holistic collections of signals, continue to provide a dual and fundamental basis for finding fundamental visual signal attributes useful in developing objective models that predict subjective responses.

I think the value of a forum like this is to suggest directions where work is needed: Going forward, there are recent innovations being made in developing perceptual models that have not yet been exploited in visual quality assessment model design. For example, recent work on the nature of responses in the cortical area V2, and how they relate to attention and visual resolution may prove to be quite valuable. Eero Simoncelli has been doing interesting work in this direction. Further along, deepening our understanding of recognition processes in inferior cortex would greatly accelerate our broader understanding of visual quality. If I were to pick an immediate problem whose resolution would have a tremendous impact on objective algorithm work, it would be the development of a consistent, regular model of the statistics of natural videos. We do not yet have a useful one that complements highly successful still image models. Whoever succeeds at developing such a model will deeply impact vision science as well as video engineering.

Moderator: Are the perceptual and objective quality measures sufficiently general to perform reliably over a very broad set of typical content? What makes the visual signal so complex to analyze and evaluate? Is it the human perception or the composition of the signal itself?

Sebastian Möller: As mentioned above, it is the internal reference that decides on quality. As an example, the same audio signal will be judged completely differently if you listen to it through a telephone connection or through your high-fidelity chain. The signal is the same, but the reference is different, and thus your quality judgment. Models for quality prediction struggle with this, as they commonly have only a marginal representation of the internal reference.

But subjective testing also reaches its limit here: If you perform an auditory quality test under narrowband conditions, the maximum rating for the clean narrowband channel will be around 4.5 (i.e., between excellent and good) on your five-point mean opinion score (MOS) scale. For the same experiment carried out under wideband conditions, the maximum rating for a clean wideband channel will also be around 4.5. Thus, network operators ask why they should introduce wideband when it results in the same MOS. The answer is that the clean narrowband channel would roughly obtain a 3.5 (i.e., a rating between fair and good) in the wideband test.

Quan Huynh-Thu: From having researched video quality prediction models for many years, I don't believe in an objective quality assessment model that does it all, either over a broad range of visual content and/or over a wide range of degradation types. Best-performing models will typically behave well for some conditions and not so well for others. On the other hand, models can gain a lot in accuracy and robustness (when their performance is compared to subjective judgment) when they are tuned for a specific context or type of content. A simple example is a model initially developed to handle both coding artifacts and transmission errors but is applied in a context to measure quality of encoded content in a head end. Obviously the model was developed to handle this scenario but the fact that it was also developed to take into account block/slice distortions due to transmission may introduce more false positives/negatives when measuring just coding artifacts. So the more general model is not always the most suitable model. Another example is videoconferencing versus video streaming. These two applications both use an audio-video signal but the context and type of content are so different that a model that was developed for video streaming is unlikely to work that well on videoconferencing, unless it was redesigned or readapted for that. In particular, the way we focus on human faces (which predominantly appear in videoconferencing) makes "talking-head" or "head-and-shoulder" content very different to other types of content.

Today, we have very sophisticated tools to compute a broad range of features to characterize/analyze the signal composition. So I believe this is not really the fundamental issue even if there is definitely still room for improvement, especially as new features may become more relevant in new scenarios. However, the integration of the information in our brain and the complex cognitive mechanisms used to come up with that single judgment of quality value are the critical points that we can't yet model robustly, especially when context and experience can have such a significant impact on the final judgment.

I think that we are currently able to predict subjective quality using computational models but clearly defining the context for which this model was developed and knowing the limitations of the model are as crucial as the performance itself.

Patrick Le Callet: I see two reasons: human perception of course and the context of use (environment, expectations, personal engagement). Both cannot be disconnected and difficult to model. It is still possible to handle part of them making right assumptions and approximations. But we are far away from a universal perceptual objective quality assessment tool, and we should remain humble regarding this challenge. Nevertheless, literature demonstrates that in some well-defined contexts, we are able to define an efficient tool. That leads to the question of reliability; one often forgets a key principle: an objective quality metric can be reliable from a perceptual point of view only if it has been validated using subjective quality assessment results. Subjective quality assessment is neither trivial, in some conditions, it might require some research efforts to design right protocols being to capture the right perceptual quality values. I am often surprised to see how some Communities that urgently need objective quality metric are keen to use new metrics violating the key principle for their proper needs. The quality assessment community should probably deliver a more cautious message regarding the good usage of objective quality metric mentioning its reliable context.

Al Bovik: Having spent most of the month of July in the galleries of Paris, Florence, and Rome, I continue to be impressed by the thought that visual aesthetics is the final ingredient of visual quality. Further, that aesthetics rely on what others here have referred to as content and context; viz., the Sunday couch football addict's aesthetic judgment is directed towards how well he or she can feel immersed in what he or she is watching, without distractions of distortions or poor camera work. How we capture the aesthetics of visual signals will be one of the great challenges of this field further out.

With all that said, and shifting my thoughts leftward, certainly some objective algorithms perform well over broad classes of content (especially full reference algorithms) but I agree with others that the experts in each domain—medical, military, mobile, and so on—will need to make these judgments.

Moderator: How successfully do the quality evaluation models emulate the human (visual, audio) perception? What are their limitations? How would you design a model if you had all the computational (and human) resources?

Sebastian Möller: In the speech quality domain, researchers have tried to model many of the known processes that are relevant for monaural auditory perception into quality prediction models. Disappointingly, some standard models [such as the Perceptual Evaluation of Speech Quality (PESQ)] have shown that an exact modeling of human auditory perception does not necessarily lead to accurate quality predictions—the models sometimes performed better if they deliberately violated human perception.

This shows that modeling human perception is not enough: Two other important aspects are missing. The first is the reference that we have already discussed. In a full-reference model, we usually take the undistorted (clean) signal as the reference, although we try to model an absolute category rating test, i.e., a test where the human listener does not have access to the clean signal for comparison. The human reference (reflecting long-term experience) might thus be different from the reference signal used in the model. This has led to proposals where a modified or idealized version of the reference is used for comparison, with some good results. The second point that is not yet appropriately modeled is the judgment process, in particular when it relates to an experience that is formed over a longer period of time. So-called "call-quality models" have addressed this point, but so far only for speech applications.

Joyce Farrell: Reference-based image quality metrics derived from models of human vision are typically designed to predict threshold judgments, such as the visibility of annoying artifacts. These metrics are relatively successful in predicting subjective judgments of image quality that are primarily driven by the visibility of noise, blur, blocking artifacts, color differences, flicker, and other types of distortions.

Reference-based metrics were not designed to predict suprathreshold judgments of image quality, such as one's preference for different types of image enhancements. Moreover, these types of judgments are influenced by cultural norms and are thus more variable across diverse populations. Therefore, it should not be surprising that metrics based on human vision models will be less successful in predicting subjective preferences.

On the question of "How would you design a model if you had all the computational (and human) resources?" I am working with colleagues at Stanford University to develop a model of the human visual system that includes human optics, photo-detector gain, retinal ganglion processing, and visual cortical processing. This is an ambitious project. My hope is that we will eventually be able to design experiments that isolate these different stages in human visual processing and determine how they influence our judgments of image quality.

Quan Huynh-Thu: Some full-reference quality models proposed in the literature and included in existing standards have shown very good performance in predicting subjective quality. However, existing models are still limited to modeling the low-level processing of the human brain, i.e., they model the visibility of the artifacts and cannot yet model robustly the (cognitive) integration process that ultimately form the subjective judgment of quality. This is in my opinion the most difficult part of the modeling because the low-level part can usually be translated into signal processing techniques, but these techniques cannot easily represent the integration part. The second limitation is the context. A model usually integrates parameters that are tuned to fit subjective databases that represent subjective quality collected in a specific context. Application of a model tuned to a context to another context is therefore not straightforward and actually should be discouraged. The context relates both to the types of degradations for which the model was designed to handle but also to the type of reference signal. Changing the reference signal will usually break the model.

Stefan Winkler: A complementary question to this would be, how much do quality models actually need to emulate human vision? Of course, in terms of outputs, we want the models to come as close as possible to subjective experimental data. However, how much actual vision modeling as such is really necessary to achieve that? Quality models using a few signal parameters and fitting functions can do surprisingly well in many circumstances.

Patrick Le Callet: I'm fully supporting this point of view. Once again, this is a matter of applications or context. From the human vision modeling perspective, applications to image and video quality assessment have been mostly limited to sensation stages, e.g., very early stages of human vision that are quite accurate to predict nearly visual threshold impairments. This is very useful in some application scenario;, the Medical Imaging Perception Society community is a good example, but when distortions deal with supravisibility threshold, this is still questionable. Applying perception and cognition concepts with a bottom-up approach is so complex than it is currently more comfortable to adopt a top-down way to proceed. This comes with some assumptions that should limit the application scope but on the other hand lead to some nice performances, as long as we remain in the scope.

Nevertheless, trying to get more generic models to be able to support more application fields is a holy quest that should help us to consider human vision much beyond than sensation stages, including higher perception level and cognition.

Stefan Winkler: To me, the question is not so much how to design a model if you had all computational resources (my tongue-in-cheek answer: build 15 or more artificial brains). The much more important question is, how much better can vision-based models actually do, and do the (often incremental) improvements over traditional approaches justify the added complexity? I believe this point is essential from a practical model usage perspective, and it is also one of the reasons the acceptance of quality models by other (e.g., video coding) research communities and engineers has been disappointingly low.

Al Bovik: Regarding images and video, I think that full reference models are doing quite a good job at predicting the quality of generic signals. Naturally there is room for improvement going forward, especially in specific application domains requiring certain types and levels of performance.

But the full reference concept is terribly limiting. To me, the Holy Grail is quality assessment (QA) models that "blindly" predict visual quality without reference, and for that matter, without knowing the distortion in advance. We and others are making good progress on this for still images. For videos, and for 3-D, the problem remains elusive since our statistical signal models are undeveloped.

If we had all the resources we needed? This is easy: quality assessment is a problem in predicting behavioral psychology using video engineering tools such as sparse and efficient representations, quantitative perception models, and machine learning. But we are lacking data. Given unlimited resources, I would conduct extremely large-scale human studies of time-dependent visual response to videos of highly diverse lengths, content, distortion types, distortion durations, and other variables. In my view, one of the great challenges is understanding how quality perception changes over time, how it relies on visual memory, and how temporal variations in quality modify quality perception over wider time scales. We will soon release a nice database and human study that explores these issues, but, alas, the videos are of usual 10–15 s presentations.

Moderator: What would be the most critical granularity for visual quality: pixel, block, frame, or sequence? What would be the similar analogy in audio quality?

Quan Huynh-Thu: Yes, this depends highly on the application and mostly what people do with the quality value. Is it for detecting a severe degradation even if it is very short in time? Is it to measure average quality on an aggregated number of communication channels? Does it need to be every 10 s or every minute? Ultimately, this depends on the usage and the level agreement between the party offering the video service and the party receiving the video service.

Stefan Winkler: As Quan already pointed out, granularity is highly application dependent. An encoder needs a very different granularity (perhaps block or even pixel-based) than the chief executive officer of a cable operator who is only interested in high-level reports (weekly or monthly). In fact, granularity is a bigger issue than people commonly realize. How do you even design a subjective test for block-level quality measurement, or for aggregating results over weeks of data, or perhaps over a number of different programs? How do short quality degradations translate into longer-term perception of quality? We know very little about how humans do these things, much less how to model it.

Patrick Le Callet: I fully agree and would like to add that is mainly due to our current lack of effort to design right subjective experiments to better understand those mechanisms. When we deal with subjective quality, usual methodologies are suitable to get short- and mid-term judgment that involves both spatial and temporal pooling. Reduced granularities are out of the game and consequently there is so far only very limited ways to construct ground truth at these levels for quality metrics.

One side additional question is: What do pixel, block, and slices mean from a human vision point of view? When designing quality metric, it is usually better to be able to translate physical parameters into perceptual space to improve robustness across various viewing conditions.

Stefan Winkler: I'd also like to raise another issue here (and my apologies for raising more questions than providing answers), on the topic of applications, or as Quan put it, what people do with the quality value.

Monitoring and comparing things is nice, but ultimately not all that exciting. If you find out something is bad, you want to know how to make it better, and by this of course I don't mean the trivial answer of "increase the bit rate" or similar. To ultimately be useful and successful, I believe quality models should be integrated in all kinds of image/video optimization and feedback loops, but very few are designed for that. Most research still focuses on the measurement aspect alone, completely disregarding the "what to do with it" issue.

This question also brings us back to granularity, not only on a temporal level, but also in terms of model output. For the type of usage I just described, having overall quality or MOS alone is insufficient, because it does not answer the question of what has gone wrong in the system, much less the question of how to fix it. Models that quantify specific artifacts go some way in addressing this, but not in a rigorous manner, in the sense that "blockiness" may be useful as another perceptual quantity, but it doesn't necessarily pinpoint a specific "culprit" either (Was it the encoder? Was it the bit rate or the content? Was it the network? Was it a loss or a delay issue?).

Joyce Farrell: I can't agree with you more Stefan. Engineers want to know how to minimize distortions and the only way they can do that is to identify the source of the distortion. As you so aptly put, they want to "fix" what has gone wrong in a system.

In the 15 years that I worked at Hewlett Packard Laboratories, I found that the people who were most vocal about the need for metrics worked in the marketing departments.

I am reminded of an anecdote someone once shared with me. Many years ago, a colleague of his was given the job to evaluate models that predict the weather. After months of research, his colleague concluded that the models they had at that time did not predict weather better than the rule "Tomorrow's weather will be like today's weather." His colleague was in the army, and when he turned in his report, he was brought before his superior, who explained to him that when commanders make a decision about a military maneuver, they cannot say that they based their decision on the theory that tomorrow's weather will be like today's weather. Luckily, in the last 50 years, we have made a lot of progress in predicting the weather. And in the future, I am hoping that our metrics will perform better than the PSNR.

But back to our discussion, I agree that engineers do not find metrics to be useful unless they help to diagnose a problem. After many years of working on image-quality evaluation in industry, I now have the time to collaborate on the design of system simulations tools (such as digital camera and display simulators) to help engineers identify the impact that different components in the image processing pipeline (from capture to processing to transmission to display) have on the final image quality. Simulation tools, in combination with metrics, can help engineers optimize the design of imaging devices.

Sebastian Möller: For speech quality, we have two projects running in Study Group 12 of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T SG12) who are dealing with a diagnostic decomposition of speech quality. One is called perceptual approaches for multidimensional analysis (P.AMD) and tries to find estimations of perceptual speech quality; the other is called technical cause analysis (P.TCA) and tries to identify the technical causes of perceptual degradations. The discussion about how many perceptual dimensions we actually need is lively, and so is the discussion about whether perceptual dimensions or technical causes are more informative—perhaps you have some ideas about this?

One of the reasons for going to perceptual dimensions instead of technical causes (at least from my point of view) was that perceptual dimensions were expected to be more stable when new transmission techniques become available. Is this the right assumption?

Al Bovik: Ultimately we should try to deploy models that resonate with perception as much as possible, but I think this can be done to a good practical approximation with any of these levels of granularity. We use them all to develop algorithms appropriate for different applications scenarios.

Ghassan AlRegib: As mentioned, this depends on applications. Humans' perception occurs at many levels so granularity and perhaps at the same time. The scope or the goal of the quality at each granularity level is different from the other levels but collectively they serve a high-level quality that is defined and required by the application. I think the correlation among the quality measures at various granularity levels is very important and needs to receive more investigation.

Moderator: What kind of errors are more disturbing: transmission errors or compression errors? How would they be balanced?

Sebastian Möller: This is the main task of a good quality prediction model, that it is able to tradeoff between different types of impairments reflecting human perception and thus can be used for taking decisions on the optimum balance. Severe transmission impairments will normally be more disturbing than compression impairments, in case that they lead to a complete loss of information, like fading in mobile speech communication, or freezing in IPTV. In addition to transmission and compression impairments, interactive services commonly also show impairments in the source material to be transmitted, such as background noise and reverberation with speech, or bad lighting conditions with video telephony. These might be even more disturbing than compression artifacts.

Joyce Farrell: I agree with Sebastian that the value of reference-based metrics is to evaluate image quality tradeoffs to find the optimal balance of different types of image impairments. Metrics based on human vision are ideal for this task because they are designed to quantify the visibility of such impairments.

Quan Huynh-Thu: More than the level of error itself, variation of error level is really annoying. If video quality is really high but video gets corrupted regularly by slice/block errors, this becomes really annoying. Conversely, if coding quality is low but remains stable, this may not be as disturbing as a video where coding quality is high but keeps dipping down regularly, due to, say, network adaptation.

Al Bovik: In simulation, any of these can be made perceptually disturbing to any degree, but in practice uncorrected transmission errors, such as packet losses, are quite annoying. As bandwidths increase, compression and resolution-related artifacts will become less significant; but as video traffic increases and wireless video continues to take hold, transmission errors in all flavors will become increasingly problematic. Keep in mind that industry predictions are for wireless video traffic to increase 5,000% or more over the next few years.

Ghassan AlRegib: Compression errors and transmission errors will continue to exist and affect the quality of received media. New types of media (e.g., 3-D video) will always require new ways to compress the content with new models and new algorithms. Also, as the bandwidth increases, the competing streams are increasing rapidly and there will always be transmission errors, especially with the rapid increase in the number of devices consumers have to access online media.

Moderator: How would social networking, collaborative tagging (in large multimedia databases such as Flickr and Picasa) and monitoring Internet browsing of content consumers change the way we evaluate the multimedia quality?

Quan Huynh-Thu: Social networking and Internet browsing are application scenarios that encompass factors that are way beyond the concept of multimedia quality as currently addressed by the quality assessment community. In those two scenarios, the notion of context, task, and interactivity are so important that the quality of the signal itself becomes only a very marginal component of the problem of quality assessment and may not be that relevant anymore as quality of experience (QoE) is what matters rather than just multimedia quality.

Patrick Le Callet: The development of such Internet services has totally modified the rules. Today, it is almost as easy to be a content producer as a content consumer. Moreover the variety of way to produce and consume has exploded, and the context of use is following. A possible consequence could be that users will start to be more educated in terms of quality requirements. Moreover, the image quality assessment community is considering the user as a watcher only while he tends to be much more than that. The new offers on interactivity (e.g., easy browsing, annotating, editing …) are some of the key factors that have participated to the emergence of consumer-producer users. I agree with Quan that QoE is what matters, as long as it is well defined and associated to a proper service and context of use.

Stefan Winkler: User-generated content is so different from professional content that we generally consider for quality assessment that it also brings with it completely different aspects about the meaning and importance of quality. Personal meaning (your family or friends are in a picture, for example) as well as timeliness (an image from a phone camera during a trip that is shared right away may be much more valuable than a high-quality picture uploaded from home three days later) often completely outweigh any quality aspects.

Joyce Farrell: The interest in visual saliency is in part driven by our desire to know what grabs consumers' attention in an image. In this sense, monitoring Internet browsing has already changed how we evaluate multimedia quality.

We need ways to evaluate the user's experience in real time. The MOS score is limited in how it captures the ups and downs of our Internet experience. Progress in both electroencephalography and functional magnetic resonance imaging make it possible to measure the user's brain activity in real time. This may become a useful evaluation tool in the future, and it is the focus of several research projects at Stanford.

Al Bovik: This is also easy. It creates a tremendous opportunity for data gathering and model development. If we are able to gather large amounts of data on human judgments of visual quality from high-traffic Internet sites (by simply asking users to supply ratings), then even in such unconstrained environments (far from the psychometric controlled viewing conditions usually demanded), the amount of data collected should prove invaluable. I think someone coined the term "social quality assessment." We are pursuing such ideas.

On the algorithm side, this is also ideal for developing learning-based visual QA algorithms. If we can expand the amount of data on visual responses to diverse contents, distortions types, severities, and temporal behaviors, then we should be able to build much more effective generic, holistic, and distortion-agnostic QA algorithms that operate without reference.

Come to think of it, this might be the answer to my "unlimited resources" question above.

Moderator: These days, each generation has different habits in the digital era. For example, young kids are very familiar with touch screens, which is not the case for older generations. Does this play a role in how quality is perceived?

Al Bovik: "Habits" is the key word here. This implies visual behavior, which as I mentioned above, is key to understanding quality perception. We will naturally find that behavioral models may vary with the interface. I think also in the "sound-bite" generation, temporal duration models may evolve and change significantly. Broadly, the expectations of consumers are for continually increased visual quality and diversity; so models will reflect this trend.

Sebastian Möller: Of course, quality does not only relate to presented media, but also to the interaction involved in obtaining and using the service. Definitely, usability plays a role here, but also the nonfunctional, or hedonic, aspects, like appeal, attractiveness, and joy-of-use. Corresponding metrics are already on the way. Unfortunately, most of the standardization bodies that deal with media quality do not work on interactive systems quality, and vice versa. The overall experience of a multimedia service will depend on both the media quality and the quality-of-use, including hedonic and pragmatic aspects.

Quan Huynh-Thu: I have previously mentioned that context is crucial in quality perception. But another factor is expectation. For a given application, the expectation is de facto related to the reference point that one has in terms of visual quality. With the evolution of the Internet technology and related services, evolution of mobile/computing devices, expectations of users change and therefore reference points also. This may not fundamentally change how visual quality could be modeled (i.e., the structure of the algorithms) but certainly changes the subjective quality benchmark. New devices also bring new ways of interaction, which can impact the way users perceive content. So far, quality assessment has mostly focused on so-called passive scenarios, i.e., where users are asked to just view/listen to content to rate its quality.

Stefan Winkler: Quality is indeed highly dependent on what people are used to and exposed to. I like to compare the times when everybody had VHS players (which were considered perfectly acceptable quality at the time) to today when people have HD TVs and Blu-ray players in their homes, and probably would cringe at the thought of watching a movie on a VHS tape.

Ghassan AlRegib: Let me target both questions above and provide one answer to both. As interactivity is increasingly becoming key in the user's experience and thus in the QoE, defining the quality of interactivity needs more investigation and research to better understand this new world. Researchers in cognitive sciences have recently figured out the "uncanned valley," i.e., why humanoid robots creep us out. The key reason is the mismatch between movement and humanoid traits. In such applications, even if the designed humanoid robot is a top technology, the lack of "realistic" interactivity movements will negatively impact the human perception. Similarly, in interacting with multimedia; if we spend all our efforts on making the multimedia quality top notch while creating an average interactivity experience, the QoE will not score high. In fact, we could use this interactivity quality to give us some space in the multimedia quality and not to demand a high media quality all the time as interactivity might compensate for the low media quality. Of course, more scientific investigation is required here to arrive at these speculations or other conclusions.

I have an App on my iPad that teaches the alphabet to my two-year old daughter. After spending some time on the iPad, she moved to the TV and tried to change the screen by touching it as she did with the iPad. To her, and perhaps, to many other kids, interacting with media is more important than the media itself. This touch-and-feel experience defines quality for them.

Joyce Farrell: Touch screens certainly affect the user's QoE. Gestures, screen size, display temporal response, and finger size are important factors that influence our experience of touch screens. It would be useful to have standardized tests to evaluate the "quality" of a touch screen, where quality is defined by task performance as well as subjective judgments. Expertise also has a very big effect. Perhaps we should study how the experience of a novice changes as he or she gains expertise in the task.

Moderator: What are the application-specific factors in multimedia quality evaluation, i.e., in 3-D displays, medical applications, online and interactive games, and streaming multimedia (e.g., speech, video)?

Joyce Farrell: This is a very good question, one that we should ask about every application. One way to answer this question is to build simulation environments that allow us to manipulate factors that we believe are important for any given application and to determine the impact they have on task performance and/or user experience.

Al Bovik: If I knew the answer to this then I would be writing proposals to quite a few different funding agencies. But seriously, the main factor is modeling. For example, a recent hot topic is the effect of visual quality on recognition (e.g., of faces). Well, to answer this, we need a face model that is useful for the problem and that is perceptually relevant. This relates to my comment on the need for better models of higher-level processing along the neutral stream. In medicine, we'll need models of the organs involved and of the physics of the imaging modality, for starters.

Ghassan AlRegib: I think such applications will affect the way we "define" multimedia quality and as a result the community will come up with ways to evaluate this newly defined quality. This is where the dimensionality and the complexity of the problem (i.e., the multimedia quality) become much harder than what we are used to. This is where the quality becomes application centric. For example, in social media, the purpose of the shared media is to convey a certain message from the individual sharing the media to her or his peers in the network; the media might be tagged or have a comment associated with it. If the message is in the audio, do we have to define the quality as a function of the audio only and not the video? If the message is in some text embedded in the video, do we pay more attention to the visual data and ignore the audio? Perhaps, if we know the "interest" of the individual, then we might have some idea on what message is being conveyed in a particular media and evaluate the quality accordingly.

Quan Huynh-Thu: I'll focus on 3-D video. In the end-to-end chain of 3-D video delivery, many points can impact quality at signal production/capture, transmission and signal rendering. Concerning the signal production/capture, several factors can impact the quality: geometric distortions/misalignment, color distortions, disparity, and imperfect 2-D-to-3-D conversion. Transmission over error-prone channels will obviously impact quality. Here we also have the headache of keeping the views of the 3-D signal synchronized in case of errors. Concerning the rendering, 3-D displays clearly play an important role in the 3-D QoE. The same stereoscopic 3-D signal displayed on two different 3DTVs may produce different experiences. There is also the adequacy between content capture and display, i.e., ideally 3-D content has to be produced for a given viewing environment (display size and viewing distance). A deviation from the target-viewing environment for which the content has been produced is susceptible to generate distortions such as shape deformation and depth distortion.

Patrick Le Callet: The impact of technology on quality of diagnosis is a seminal issue while considering medical imaging systems such as magnetic resonance imaging and a positron emission tomography scan. The quality of a system relies on its ability to minimize wrong decision of the practicians. This is an extreme case that required full adaptation to the application scenario. Signals are very specific (sometimes the relevant information looks more like noise for nonexperts) and to assess their quality it is needed to consider the pathology under study, the related anatomy, and the image modality. This is part of the medical imaging readers' expertise that a metric should be able to mimic. Of course, it is more realistic to tune a metric for one particular combination of pathology/anatomy/image modality. As all the elements of the chain, from acquisition to visualization, might affect the final decision, it is required to well understand the consequence of technological aspects on the signal while measuring the value of one particular element of this chain.

Moderator: 3-D video quality is an emerging field; what are the challenges ahead?

Al Bovik: What a booger this problem is! First, we lack models of the statistics of the natural 3-D world. We will require this to make good progress. Second, despite 40 years of research of stereopsis and other modes of 3-D, we lack understanding of key perceptual elements of 3-D perception. For example, we do not understand yet how the stereo sense affects the perception of distortions. Does high stereo disparity activity make luminance distortions less visible similar to luminance masking? Recent studies suggest, unexpectedly, that the opposite might be true. Does viewing in 3-D change where we look? Certainly, and unexpectedly, it seems. There are many other examples where our understanding is quite poor. This compounds the fact that 3-D video quality is a double-blind problem: since the perception of the 3-D signal (distorted or otherwise), termed cyclopean signal, occurs only in the brain, we have no way to quantitatively access either the distorted signal or a reference signal. We are thus left with estimating these, before predicting quality. As others have pointed out, 3-D QoE is multifaceted with discomfort, distortion, and display issues (one might call these the three "D"s) all playing a role. However, we poorly understand how each of these relate to overall QoE, and we haven't begun to understand how they combine to affect QoE. We are taking a step-wise approach by examining these issues separately. In particular, I strongly disagree with one of our panelists regarding the success of 3-D quality (distortion) models. We haven't found any that improve upon 2-D models as applied to 3-D data. This remains an open area of research.

Ghassan AlRegib: 3-D video brings interesting challenges to the community on how to evaluate the quality of 3-D videos. Here we are looking at a number of monocular cues from two or more views, and the display is trying to recreate the 3-D world. All steps in the pipeline from acquisition to coding and from streaming to displaying affect the quality of the 3-D video. The impact is usually catastrophic and it results in a great deal of discomfort.

If we consider stereo videos, for example, a slight variation in the color between the two views will result in a discomfort. In depth-based videos, we overcome this problem by using a single reference to generate the second view but this brings it own challenges such as occlusion and disocclusion.

Another challenge is the fact that we do not have a reference in 3-D videos. We capture imperfect views and we try to reproduce 3-D video that has the least artifacts and the most natural views even though the captured views we started with are far from perfect. This introduces challenges, but it also leaves room for us to innovate in processing and displaying the 3-D videos.

From our research at Georgia Tech, we found out that it is a failure to think of the 3-D video as the combination of two 2-D signals. In this case, a few quality measures have been proposed where the overall quality measure is the combination of individual views quality; for example, the average PSNR of the two views. This oversimplification results in mismatch between measured quality and subject quality. In contrast, one has to consider a 3-D video as a 3-D signal and create a quality measure that is designed for a 3-D signal with certain components.

Finally, the whole R-D analysis for 3-D videos is based on a set of "new" quality metrics that are needed to be designed for 3-D videos. This will open the door for a number of innovative approaches and algorithms.

Quan Huynh-Thu: The concept of 3-D video quality is in fact multidimensional: it includes the signal quality but also other dimensions such as visual comfort and depth quality. These could be termed basic perceptual dimensions. Existing 2-D video quality assessment algorithms are designed to address only the first one (signal quality). 3-D video quality is not simply an extension of 2-D video quality with depth information. Furthermore, even if we consider only the signal quality, there are artifacts that are only specific to 3-D video and not existing in 2-D, so existing 2-D quality assessment algorithms are not designed to handle such 3-D-specific distortions. Second, the subjective assessment of all these dimensions, in particular visual comfort and depth quality, is not easy. Keep in mind that subjective assessment is only meaningful if results are repeatable and reproducible. Currently, it is not clear how to reliably and meaningfully assess subjectively visual comfort and depth sensation/quality of stereoscopic 3-D motion sequences, especially for long durations. Third, in 3-D video, not only the signal itself but also the rendering of the signal can impact significantly the subjective quality. 3-D displays have improved over the years but still suffer from crosstalk, which can decrease the 3-D QoE. Last but not least, in the case of 3-D video, the true signal to assess should in fact be the one reconstructed inside the brain and not the signal displayed on the screen. The bottom line is, if one wants to assess the quality of the 3-D signal as truly perceived, the binocular fusion process should somehow be modeled and integrated—this is not trivial as we need to understand how we actually integrate all the different (monocular and binocular) content cues together.

A paper published at the 2010 IEEE International Conference on Image Processing titled "Video Quality Assessment: From 2-D to 3-D–Challenges and Future Trends" summarizes why 3-D video quality metrics cannot be simple extensions of 2-D video quality metrics and which points in the 3DTV transmission chain are susceptible to affect the QoE at the end-user side.

Patrick Le Callet: As a coauthor with Quan of this paper, I of course fully agree with this vision, with two more comments:

- Lack of ad hoc ways to measure subjectively 3-D QoE is currently a trap for most of quality metric designers. Using standard 2-D quality assessment protocols leads to oversimplification of ongoing 3-D quality metric efforts as pointed out by Ghassan. There are a bunch of 3-D quality metrics in literature that are quite successful to correlate with MOS obtained using standard protocols catching the visual quality. This is just one piece of the puzzle, as other key perceptual dimensions like discomfort and depth sensation are out of this game. I definitively recommend being very cautious while interpreting the results of such metrics. For instance, there is a trend to use asymmetric coding conditions to save bandwidth for one of the view in stereo transmission. In terms of visual quality, it might work as visual perception of most observers is able to align the quality on the "best" view. Nevertheless, this compensation might fire a cognitive load that could have some dramatic effects considering the discomfort dimension. This effect will be transparent for most of usual subjective quality assessment methodologies that uses short video clips, and consequently also for quality metrics tuned on such ground truth.
- 3-D right now means mostly stereo 3-D (S-3-D) (providing two different views to the use whatever the display, shutter, passive, and lens auto stereoscopic displays). It cheats our visual perception enhancing one depth cues: binocular disparities, but we are able to perceive depth from many other cues. Enhancing one of the cues might affect the other ones in a nonreliable way (e.g., losing resolutions, due to interleaved stereo format, affect monocular cues such as texture gradients) that leads to the question of the value of such solution. Considering the challenge ahead, quality assessment should not only focus on S-3-D but should be provide some answers on the values of all candidates technologies such as motion parallax-based ones, holography, etc.

Stefan Winkler: As others have already elaborated, 3-D quality is a highly complex subject with many different issues, which we have only begun to explore and understand.

I'd like to highlight another important aspect: In 3-D it's not just about the best-looking content anymore. Stereoscopic content has potential psychological and physiological effects on a significant portion of the population: if 3-D is not produced, processed, and presented correctly, it can make viewers dizzy or nauseous. Understanding why this happens for some people more than for others, and how these effects can be minimized will be crucial for the success and wide adoption of 3-D.

Joyce Farrell: There are challenges in nearly every aspect of 3-D imaging, including image capture, processing, transmission, and display. But the biggest challenge, it seems to me, is to quantify the value of 3-D when it comes to the users' experience. Industry's fascination with 3-D displays resurfaces every ten years or so. Will this be the year that 3-D displays find their way into people's homes and theatres? For me, this is still an open-ended question. Even if we can solve the technical challenges, will people value the 3-D experience enough to pay for it?

There are at least two applications for which the answer to this question is yes. One application is medical imaging, such as robotic assists in products like the

DaVinci surgical unit. Another application is video games. The impact of 3-D imagery on these applications is undisputable. But for someone who is not a surgeon and who does not play video games, the question of how much I will pay for 3-D imagery remains unanswered.

Moderator: How might the scientific progress and discovered principles in multimedia quality evaluation research benefit other fields?

Quan Huynh-Thu: Research on multimedia quality, and on objective metrics in particular that can predict subjective quality, is a crossroad between several fields such as image processing, computer vision, cognition, neurosciences, and psychology. So the understanding and modeling of how we perceive a signal and form an opinion of quality ultimately benefit the knowledge in all these fields.

Patrick Le Callet: It helps to better tune the technology to the final user. Whatever the multimedia application, adopting a user-centered approach for the development of technologies as a piece of a system of a whole product is much welcomed as it is done for the design of end-user products. The user-centered product design community has certainly developed excellent approaches nevertheless face some difficulties while addressing the lower-level pieces of technologies. For the latter, a good understanding of the underlying technology is mandatory in addition to human factors considerations. This is where our community could be helpful: to fill the gap between engineers, cognitians, and designers.

Stefan Winkler: I don't think we need to look very far for other fields where quality assessment can be beneficial.

As I mentioned earlier, good quality models should be integrated in all kinds of multimedia processing chains for best results, but very few are designed for that. Better quality metrics than PSNR should be used in encoders for rate control, in LCD displays for content preprocessing, in cameras for photo optimization, etc. Unless the quality assessment community can come up with metrics that not only perform well but are also easy to use, easy to interpret, and well accepted, adoption by other communities will be even slower.

Joyce Farrell: Perhaps we should ask how can we make scientific progress and discover principles in multimedia quality evaluation. One approach that I advocate is to build simulation environments that control every aspect of the multimedia signal, including properties of the capture, processing, transmission, and display. In this way, we will be able to determine the critical components that influence subjective quality. Of course, we still have the question about how best to characterize the users' experience.

Clearly, evaluating the user's experience is a challenge. Today, we ask people about their experience after it has already happened. We are initiating a research project at Stanford that will monitor users' brain activity in real time as they view video imagery, both 2-D and 3-D. Whether this produces better information than subjective reports about fatigue or enjoyment is an open-ended but important question to answer.

Al Bovik: Aside from the seeming limitless realm of image/video/multimedia applications that will benefit by using quality models to monitor, assess, and control the quality of visual signal traffic using successful QA models, in my opinion, understanding visual quality and visual distortion perception is fundamental to understanding vision. I view visual distortions and how they are perceived as visual probes into perception, much as visual illusions are. If we come to understand how distortions and their severity are perceived, then we will likely have made significant inroads into understanding a wealth of other visual principles.

Ghassan AlRegib: I think understanding how to evaluate the quality of multimedia will help us understand the source of such complex signals, e.g., human speech, light fields, human vision, etc. This will open the door for many scientific discoveries. For example, if we truly know how the human brain and vision system views things in terms of quality, then we can design better car "visual" systems and better robots.

Moderator: In a recent standardization effort, specification of the hardware the displays was well spelled out. What role does hardware play in affecting how visual quality is perceived? Do we need to come up with a set of quality measures that depend on the hardware?

Sebastian Möller: I think that we should always take into account the full channel, including the sending and receiving (hardware and software) elements. A model that does not explicitly mention the hardware used will make, nonetheless, assumptions about it, specifically from the test situations that have been used for collecting the data material for the model. These assumptions should be made explicit, and care should be taken when interpreting the model predictions in cases where the sending and receiving elements are different.

Patrick Le Callet: Displays are the ultimate steps that translate information into the real world. As long as we are considering mature enough technology, this step can be modeled once for all and being transparent for some metrics. Regarding S-3-D jungle technologies, not only displays but also formats better be cautious and properly define the conditions of the validation of a metric.

For peaky applications such as medical imaging, quality of the display, from a technological point of view, is a strong issue.

There are several standardization efforts to define quality of displays [International Committee for Display Metrology (ICDM), American Association of Physicists in Medicine (AAPM)], and most of them are providing very useful measurements of the technological parameters of displays, sometimes considering human perceptual properties. Image quality assessment, as going toward QoE and targeting more application specific context, should better try to understand the impact of these quality displays factors on the overall visual perceptual quality.

Quan Huynh-Thu: In any standard, the scope is highly important. A technology such as a quality assessment algorithm is designed for a given scope (e.g., types and severity of degradations) and has been validated on subjective data collected in a given scenario, which includes a certain type of display (specifications). In recent standardization efforts by the ITU supported by extensive work from the Video Quality Experts Group (VQEG) concerning (2-D) video quality metrics (ITU-T J.247, ITU-T J.341), studies have examined whether the display had a significant impact on the subjective quality. It was found that for the given scope of degradations considered in the subjective testing, a very high similarity was found between subjective results even if different displays were used, provided the displays had some minimum performance criteria. So ultimately the standard does not specify any display for which the metric has been validated but de facto a quality assessment metric has been validated on subjective data that has been collected in a certain scenario (including the display that was used to show the videos to the participants). With the current state-of-the-art 2-D display technology, the influence of the display on visual quality is highly dependent on the application. For the types of applications considered in ITU-T J.247 and ITU-T J.341, the display (again with some minimum criteria) did not impact the visual quality (especially for naive viewers) but in other applications, where, for example, color fidelity is crucial or in medical imaging, obviously display is important. Now that said, 3-D video is a different case, as I commented previously. The current 3-D display technology has not reached a level of maturity where the display can be considered to be transparent and clearly there is a need to either define metrics that include the influence of the display or make sure that the display used in subjective testing is "transparent" enough.

Stefan Winkler: Displays clearly play an important role in determining QoE. Perhaps the wider question is, how much different is the experience in an actual home viewing environment from subjective experiments performed in the lab under rigorously controlled conditions? This is not only about the quality and settings of the display, but also about viewing setup, lighting conditions, length of viewing (a whole movie versus a 10-s clip), attention, interest, etc. Of course, we can design quality metrics to work under the "best possible" viewing conditions and displays, but it may be worth exploring how these other parameters can be taken into account.

Joyce Farrell: Quality depends on hardware, software, and the properties of the measurement device, be it a human or an instrument.

At a minimum, we should report the conditions in which we collect subjective judgments of image quality, including the viewing distance, the ambient illumination and the spectral power of the display primaries, the display gamma, and the display resolution. We should also be sure that subjects have corrected to normal vision. And it may be useful to record their age and sex. I hope that in the future, databases that match images and video with mean opinion scores (MOS) will include this information.

And if we want to understand the role that hardware (and software) plays in determining subjective image quality, we need to be able to independently control different hardware (and software) components and record their impact on subjective judgments of image quality. This is why I am a strong proponent of simulation environments.

Moderator: What is the next challenge in the quality arena?

Sebastian Möller: One of the challenges I see is to come to realist predictions of what people do, i.e., how they behave when being confronted with a service. We have invested a lot in making predictions about people's perceptions, but not so much about people's actions. However, when we deal with interactive services, the user's actions will be of paramount importance for the overall quality. What is needed are models that describe user actions both on a semantic (intentional) and on a surface level, i.e., the level of observable actions.

Another challenge I see is to link quality to other aspects affecting the acceptance of a service. One of them is the price or economical benefit. Another is security. People inherently establish tradeoffs between these aspects, and it would be wise to not concentrate on quality alone when designing a new service.

Patrick Le Callet: It appears that with the explosion of applications and technologies, QoE assessment is the key. This is not a new concept: I still remember Touradj Ebrahimi mentioning this emergency in 2001 during a conference keynote, ten years ago already, but we have still far to go to provide satisfying solutions. How do we go beyond? Following

Perreira's triple sensation-perception-emotion user model is a way that tries to measure the following different steps: 1) sensation factors that relate to the sensory perception of (multimodal) content, i.e., human vision, audition, and other sensory perception; 2) perception factors that relate to interpretation of the information from the media, user's satisfaction as cognitive experience (usability, task performance, and information assimilation); and 3) emotion factors that relate to the intensity of emotional experience.

Another way could be to see QoE from a user's experience following the Mahlke model of user experience. As components of user experience, Maklke identifies: 1) the perception of instrumental qualities: usefulness (utility), usability (efficiency, controllability, helpfulness, and learnability); 2) emotional user reactions: subjective feelings, motor expressions, physiological reactions, cognitive appraisals, and behavioral tendencies; and 3) perception of noninstrumental qualities: aesthetic aspects, visual aesthetic, symbolic aspects, associative symbolic, communicative symbolic, motivational aspects, and multimodalities.

Whatever the approaches, QoE assessment will probably occupy researchers for many years. Hopefully, there are some ongoing collaborative actions towards this goal. The European Cooperation in Science and Technology (COST) action network on QoE in multimedia systems and services (QUALINET) is currently gathering the efforts of partners from more than 30 countries toward this goal.

As short-term realistic goals, I see two interesting scopes beside 3-D for quality assessment. The first one is dealing with immersion through two technological aspects: super high resolution and high dynamic range (HDR). The second one is related to multimodality, trying to better understand relationship between visual and haptic perception.

Ghassan AlRegib: In the short term, I think the overall immersive technologies will be heavily investigated. This includes 3-D and interactivity via haptics or via touch. Also, social media will receive quite a bit of attention to determine the quality of a social network based on communicated media and the associated reactions and responses.

Overall, I agree with Patrick, and I believe we are still far from having clear and thorough understanding of how perception, sensation, and emotion interact to define "quality." Add to this the complex pipeline of systems media content undergoes before it is perceived by the user.

Quan Huynh-Thu: Quality measurement alone is not really meaningful per se. In a real-world scenario, its meaning is always linked to a context, which can include the application/service and fees that users must pay. Quality measurement has been so far used mostly for troubleshooting and network/application tuning. Other points that are related to quality are usability and acceptability. Linking these two points to quality measurement is not easy and so application/service dependent that finding a generic way to model their relationship would be quite remarkable. We would almost need the types of models they use in finance.

Stefan Winkler: I see three main challenges: One is new technologies, such as 3DTV, or upcoming display technologies. Another is related to my comments on granularity: We need to find useful applications for quality measurement beyond just monitoring and comparisons, and for that we need to look beyond MOS and overall quality. The third is the rather narrow signal processing focus we usually have in terms of what quality constitutes. Aspects of interaction, ease of use, personal preference, context, emotion, relevance, and appeal, are as important (if not more) than compression and other distortions and need to be taken into account for a comprehensive assessment of the true "QoE."

Joyce Farrell: One challenge is to develop new methods and standards for assessing the user's experience. MOS is clearly limited. Rather than one method, we need multiple methods that quantify different aspects of the quality of multimedia.

Another challenge is to control the complex interactions between multiple factors that influence the user's experience. Video content, image capture, processing, transmission, display, viewing conditions, and user characteristics (expertise, age, sex, vision, etc.) all influence the quality of the multimedia experience. We need ways in which we can independently control different components of a multimedia system as well as methods for analyzing design tradeoffs.

Al Bovik: I believe that as our models improve and become adaptive and intelligent, we should seek to deploy "visual quality agents" in every switch, router, access point, TV, smart phone, camera, and so on. These quality agents should ultimately interact, enabling distributed network control of video traffic, and perceptually optimized acquisition, transmission, coding, compression, and display of visual information. This implies huge deployments, and yes, I think the problems we are working on are this important.

MODERATOR

Fatih Porikli (fatih@merl.com) is a Distinguished Scientist at Mitsubishi Electric Research Labs (MERL), Cambridge, Massachusetts, United States. He received his Ph.D. degree from the Polytechnic Institute of New York University. Before joining MERL in 2000, he developed satellite imaging solutions at Hughes Research Labs and 3-D capture and display systems at AT&T Research Labs. His work covers areas including computer vision, machine learning, compressive sensing, sparse reconstruction, video surveillance, multimedia denoising, biomedical vision, radar signal processing, and online learning. He received the 2006 R&D100 Award in the "Scientist of the Year" category (a select group of winners) in addition to numerous best paper and professional awards. He serves as an associate editor for many IEEE, Springer, and SIAM journals. He was the general chair of the 2010 IEEE International Conference on Advanced Video and Signal-Based Surveillance and organizer of several other IEEE conferences.

PANELISTS

Al Bovik (bovik@ece.utexas.edu) holds the Curry/Cullen Trust Endowed Chair professorship at The University of Texas

at Austin, where he directs the Laboratory for Image and Video Engineering (LIVE) in the Department of Electrical and Computer Engineering and the Institute for Neurosciences. He is broadly interested in image processing and modeling of visual perception. He has received many awards for his work including most recently the 2011 IS&T Imaging Scientist of the Year Award and the 2009 IEEE Signal Processing Society Best Paper Award. He is also known for helping to create the IEEE International Conference on Image Processing (ICIP) and cofounding *IEEE Transactions on Image Processing*.

Chris Plack (Chris.Plack@manchester.ac.uk) is the Ellis Llwyd Jones Professor of Audiology at the University of Manchester. He specializes in human auditory perception, in particular, the physiological mechanisms that underlie our perception of loudness and pitch, and the effects of hearing loss on these mechanisms. He is a fellow of the Acoustical Society of America, and an associate editor of *Journal of the Acoustical Society of America*. He is the author of *The Sense of Hearing* (Psychology Press) and has edited two volumes: *Pitch: Neural Coding and Perception* (Springer) and *Hearing* (Oxford).

Ghassan AlRegib (alregib@gatech.edu) is an associate professor at the School of Electrical and Computer Engineering at the Georgia Institute of Technology in Atlanta. He is the director of the Multimedia and Sensors Lab at Georgia Tech. He received the 2008 Outstanding Junior Faculty Award. He is conducting research in the area of multimedia processing with recent focus on 3-D videos quality, processing, and compression. He is the area editor for columns and forums in *IEEE Signal Processing Magazine* and is the editor-in-chief of *ICST Transactions on Immersive Telecommunications*. He is the cochair of the IEEE Multimedia Communications Technical Committee Interest Group on 3-D rendering, processing, and communications. He is an IEEE Senior Member.

Joyce Farrell (joyce_farrell@stanford.edu) is a senior research associate in the Department of Electrical Engineering at Stanford University. She is also the executive director of the Stanford Center for Image Systems Engineering. Prior to joining Stanford University, she worked at a variety of companies and institutions, including the NASA Ames Research Center, New York University, the Xerox Palo Alto Research Center, Hewlett Packard Laboratories, and Shutterfly. She is also the chief executive officer and founder of ImagEval Consulting, LLC.

Patrick Le Callet (patrick.lecallet@univ-nantes.fr) is professor at Polytech Nantes-Université de Nantes and head of the Image and Video Communication group at CNRS IRCCyN. He is mostly engaged in research dealing with human vision modeling and its application in image and video processing with the current center of interest in color and 3-D image perception, visual attention modeling, video, and 3-D quality assessment. He serves in the VQEG where he cochairs the Joint Effort Group and 3DTV activities. He is the French national representative of the European COST action IC1003 QUALINET on QoE of multimedia service in which he leads the working group on mechanisms of human perception.

Quan Huynh-Thu (qht@ieee.org) is currently a senior scientist at Technicolor Research & Innovation. His main research interests for the past ten years have been video quality assessment, human factors, and visual attention. He holds the Dipl.-Ing. degree in electrical engineering from the University of Liège (Belgium), the M.Eng. degree in electronics engineering from the University of Electroc-Communications (Japan), and the Ph.D. degree in electronic systems engineering from the University of Essex (United Kingdom). He codeveloped a perceptual video quality metric included in the ITU-T Recommendation J.247 for the objective measurement of video quality. He is a rapporteur for Question 2 in ITU-T SG9 and cochair of both the VQEG 3DTV and multimedia groups.

Sebastian Möller (sebastian.moeller@telekom.de) studied electrical engineering in Bochum (Germany), Orléans (France), and Bologna (Italy). He received his Ph.D. degree from Ruhr-Universität Bochum in 1999, and his qualification to become a professor (venia legendi) in 2003. Since 2007, he has been a professor for quality and usability at Deutsche Telekom Labs, TU Berlin, and works on speech and audio-visual quality, spoken dialogue systems, usability, user modeling, and usable security. He is currently acting as a rapporteur for ITU-T Q.8/12.

Stefan Winkler (stefan.winkler@adsc.com.sg) is a principal scientist at the Advanced Digital Sciences Center of the University of Illinois at Urbana Champaign in Singapore. He is also a scientific advisor to Cheetah Technologies. He holds a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, and an M.Eng./B.Eng. degree from the Technische Universitat Wien, Austria. He has published more than 70 papers and is the author of *Digital Video Quality* (Wiley). [SP]